Deep Learning-based Audio Source Separation, Recognition, and Information Extraction for Multimedia Analysis

著者	高橋 直也
発行年	2020
その他のタイトル	マルチメディア理解のための深層学習を用いた音源分離、認識、及び情報抽出に関する研究
学位授与大学	筑波大学 (University of Tsukuba)
学位授与年度	2019
報告番号	12102甲第9406号
URL	http://hdl.handle.net/2241/00160955

氏　名	高橋 直也
学位の種類	博　士（工学）
学位記番号	博　甲　第　９４０６　号
学位授与年月日	令和２年３月２５日
学位授与の要件	学位規則第４条第１項該当
審査研究科	システム情報工学研究科
学位論文題目	Deep Learning-based Audio Source Separation, Recognition, and Information Extraction for Multimedia Analysis （マルチメディア理解のための深層学習を用いた音源分離、認識、及び情報抽出に関する研究）

主　査	筑波大学　教授	博士（工学）　牧野　昭二
副　査	筑波大学　教授	博士（工学）　福井　和広
副　査	筑波大学　教授	博士（工学）　亀山　啓輔
副　査	首都大学東京　教授	博士（工学）　小野　順貴
副　査	筑波大学　准教授	博士（工学）　山田　武志

論　文　の　要　旨

大量のマルチメディアがアクセス可能になるにつれ、コンテンツからの情報抽出やサマライズ、コンテンツに対する質問への応答などのタスクのための機械によるマルチメディア理解が益々重要になっている。現状この様なタスクは人手で作成されたラベルに大きく依存して行われているが、その様なラベルの作成は非常に時間がかかったり、不正確であったりするため、自動化するためには機械がマルチメディア自体を解析することが重要である。近年の深層学習の発達により機械によるマルチメディア解析は進歩を見せているが、多くの場合、依然としてそのような技術は大量のデータが利用可能なドメインに限られており、解析はコントロールされた、又は限られた環境でのみ行われている。例えば、多くの音の認識技術は大量のデータが利用可能な主要な言語の音声認識に集中しており、音声は単一話者のクリーンな環境下のものであることが多い。本研究ではより現実的な、コントロールさていない環境での機械によるマルチメディア理解を実現するにあたって重要な問題を、特にオーディオドメインについて扱っている。具体的には複数音源の混合音に対する問題、データセットのリソースが限られている場合の問題、そして非音声信号を含む音情報をビデオ解析に利用する方法について扱い、新たな音源分離手法、低リソース言語の音声認識手法、音響イベント認識手法、行動認識やハイライト検出を含む音画像ビデオ解析手法について提案し、実験により評価し、有効性を確認している。

審　査　の　要　旨

【批評】

本研究では，現実的な、コントロールさていない環境での機械によるマルチメディア理解を実現するにあたって重要な問題を、特にオーディオドメインについて扱っている。具体的には

1) 複数音源の混合音に対する音源分離のための効率的なネットワーク構造として MMDenseLSTM を提案している。このアーキテクチャは DenseNet を応用したもので、スペクトログラムを畳み込みと再帰構造の両方を備えるネットワークを用いて多重解像度、多重バンドで モデル化している。提案手法は音楽分離のコンテスト SiSEC2018 でベストスコアを獲得した。また、音源分離における目的音源の位相を推定する問題において、位相を離散化し、DNN で分類問題として推定する手法を提案している。話者分離問題では混合音内の話者数が未知の場合に単一モデルですべての話者を再帰的に分離する手法を提案し、実験により評価し、有効性を確認している。

2) データセットのリソースが限られている低リソース言語のための音声認識では、データからサブワードと辞書を同時に学習する手法を提案している。単語の区間のみが既知の状態から弱教師あり学習で音響モデルと辞書を交互に更新し、精度を徐々に上げていくことで過学習を抑えた高精度なモデルを学習している。実験により提案手法は人手で設計された音素と辞書を用いて学習したモデルを上回る音声認識精度が得られることを示している。

3) 音響イベント認識では、非音声を含む一般的な環境音を DNN を用いて認識する手法を提案している。具体的には、スペクトログラムの長時間フレームを一括で DNN でモデル化、小さいカーネルと深い層を持つネットワーク構造、新たなデータオーグメント手法を提案し、実験により評価し、有効性を確認している。

4) 行動認識やハイライト検出を含む音画像ビデオ解析手法では、ビデオ解析のための音響特徴量として転移学習を用いた手法を提案している。 41 クラスの音響イベントデータセットを作成し、 音響イベント認識 タスクを学習した DNN の中間層の出力を特徴量として用いる。行動認識、及びビデオハイライト検出タスクにおいて、画像特徴量と既存音響特徴量を用いた場合と比べて大きな性能向上が得られることを示し、提案特徴量の有効性を確認している。

研究の着眼点、新規性、有効性、実用性において極めて優れた研究であり、博士（工学）の学位を受けるにふさわしい優れた論文と評価する。

【最終試験の結果】

令和２年２月５日、システム情報工学研究科において、学位論文審査委員の全員出席のもと、著者に論文について説明を求め、関連事項につき質疑応答を行った。その結果、学位論文審査委員全員によって、合格と判定された。

【結論】

上記の学位論文審査ならびに最終試験の結果に基づき、著者は博士（工学）の学位を受けるに十分な資格を有するものと認める。